POWER, PERFORMANCE, AND COST IMPACT OF GATE-LEVEL MONOLITHIC 3D IC IN THE 7NM TECHNOLOGY NODE

A Dissertation
Presented to
The Academic Faculty

By

Bon Woong Ku

In Partial Fulfillment
of the Requirements for the Degree
Master of Science in the
School of Electrical and Computer Engineering

Georgia Institute of Technology

May 2017

POWER, PERFORMANCE, AND COST IMPACT OF GATE-LEVEL MONOLITHIC 3D IC IN THE 7NM TECHNOLOGY NODE

Approved by:

Dr. Sung Kyu Lim, Advisor
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Saibal Mukhopadhyay
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Arijit Raychowdhury
School of Electrical and Computer Engineering
Georgia Institute of Technology

Date Approved: April 27, 2017

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

The nature of device scaling in the 7nm era is changing from the traditional scheme driven by Moore's law because of the physical and economic limitations of fabrication. Instead of reducing a horizontal distance between devices to increase the density of system functionalities, concerted efforts for 3D integration have been made from both industries and academic fields to utilize the vertical dimension to increase the chip density.

Over the last few years, monolithic 3D (M3D) technology, which involves the integration of one or more active layers on top of a prefabricated metal stack in monolithic fashion, has emerged as a promising solution for the massive 3D interconnection. The objective of this research is to study the impact of M3D technology on the power, performance and cost of integrated circuits under unique challenges of M3D technology in the 7nm node.

The most critical challenge in M3D integration is that once the bottom tier devices and interconnects are implemented with the normal process, they suffer from additional thermal exposure during the dopant activation step of the top tier. Therefore, alternative fabrication steps and materials for each tier are required. The first section of this thesis presents the physical design methodologies to tackle the inter-tier variations in 2-tier M3D ICs, and shows the power and performance benefits of M3D ICs in various scenarios.

Although M3D integration offers the small form factor, low power, and high system performance, high fabrication cost is another challenge to justify the adoption of M3D technology. The second section investigates the complicated power, performance, and cost (PPC) tradeoffs of 2-tier M3D ICs based on the accurate wafer and die cost models along with the optimal CAD solution for M3D ICs to maximize the area and routing utilization of designs.

CHAPTER 1
INTRODUCTION

1.1 Monolithic 3D (M3D) Integration

Active devices have improved its switching power, delay and size for the last fifty years based on the geometric scaling driven by Moore's law. However, the nature of device scaling in the 7nm era is changing. Although the bulk FinFET device is still adopted for the 7nm technology node according to the announcement of chipmakers, both industry and academic fields are exploring 3D device architectures such as the nanowire and the vertical FET for the following technology nodes [1]. A main reason that device scaling is enforcing its 3D nature is because of the high fabrication cost to realize the geometric scaling in the advanced technology nodes. An increase in the number of lithography masks to enable the small feature size of the 7nm node seriously diminishes the economic benefit [2]. Power-performance benefits of geometric scaling also decrease because of the subthreshold degradation and internal parasitics in the advanced nodes [3]. Moreover, a delay in introduction of extreme ultraviolet lithography technology to the manufacturing significantly affects the approaches to the next generation device technology [4, 5].

Over the last few years, monolithic 3D (M3D) technology, which involves the integration of one or more active layers on top of a prefabricated metal stack in monolithic fashion, has emerged as a promising solution to overcome the physical and economic limitations of logic scaling in the 7nm node [6, 7]. While traditional through-silicon-via (TSV)-based 3D integration requires a die alignment step after dies are fabricated in parallel, M3D integration fabricates the dies sequentially retaining the litho-scale alignment precision. Therefore, while the size of a TSV should have its μm-scale lower bound to avoid unexpected disconnection during the die alignment process, monolithic inter-tier vias (MIVs) achieve

the tier connections in M3D integration at *nm*-scale. The extremely small size of MIVs not only minimizes the area overhead of the vertical connection, but also offers the number of inter-tier connections in orders of magnitude. Therefore, effectively inserted MIVs significantly reduce the wirelength of 3D nets, resulting in maximized power-performance benefits of 3D integrated circuits (ICs) even more than the benefits from logic scaling [8].

Depending on the granularity of MIV insertion, M3D integration is categorized into the transistor-level (T-M3D), gate-level (G-M3D), and block-level (B-M3D) [10, 11, 12]. While T-M3D uses ultra-dense MIVs inside standard cell designs to connect transistors placed on separate tiers, G-M3D and B-M3D use MIVs to route the 3D nets of blocks or gates placed on multiple tiers. Compared to B-M3D, G-M3D uses more dense vertical interconnections, resulting in the sufficient wirelength savings in the global routing. In addition, G-M3D allows to reuse existing 2D standard cell libraries for the physical design of M3D ICs, while T-M3D requires a new layout and characterization of standard cells. Therefore, this research studies the power, performance, and cost (PPC) tradeoffs for 2-tier, full-chip G-M3D ICs built on a foundry-grade 7nm bulk FinFET technology under the unique challenges of M3D integration.

1.2 Unique Challenges of M3D in the 7nm Technology Node

1.2.1 Limited Thermal Budget

One of the most critical problems in M3D integration is the limited thermal budget for the top tier fabrication process. Once the bottom tier devices are implemented with the normal process, they suffer from additional thermal exposure during dopant activation step of the top tier ($T > 1000\,^{\circ}C$). In the meantime, integrating Copper (Cu) interconnects in the bottom tier implies that thermal budget for the top tier has to be under $450\,^{\circ}C$, because Cu diffuses away into the Low-K regions at such a high temperature. In order to preserve the device performance and the integrity of the back end of line (BEOL) of the bottom tier, recent studies [13, 14] introduced molecular wafer bonding and solid phase epitaxial

regrowth (SPER) dopant activation process for the top tier device manufacturing based on planar FDSOI device. In the industrial environment for the 7nm technology node, however, implementing FinFET or nanowire devices requires conformal in-situ doping in S/D region due to the 3D structure of the device, leading to high temperature annealing processes (T $> 1100\,^{\circ}$C). Therefore, the limited thermal budget of the top tier fabrication is expected to bring serious performance loss on the device. Tungsten (W) interconnects of the bottom tier is an alternative to offer more thermal budget for the top tier manufacturing, but the high resistivity of W degrades the overall performance.

1.2.2 High Fabrication Cost

The high fabrication cost is another challenge to justify the adoption of M3D technology in the 7nm technology node. Since M3D integration requires both front end of line (FEOL) and BEOL for each tier, high M3D wafer cost degrades the die cost savings from the small design footprint of M3D ICs. Moreover, both FEOL and BEOL fabrication cost has been increasing as the dimensional scaling advances toward aggressive pitches because of the increasing number of photomasks for multiple patterning. Even though introduction of extreme ultraviolet lithography is expected to reduce manufacturing cost, complicated device structure and the growth of system design complexity in the 7nm node degrade the die yield [5, 2]. To reduce the M3D wafer cost, industry must reduce the number of metal layers of each tier to the maximum allowable extent, but the routing congestion from the limited metal resources makes the complicated power, performance, and cost (PPC) tradeoffs in M3D ICs.

1.3 Organization and Contributions

This dissertation first presents the physical design methodologies to tackle the inter-tier variations in two-tier G-M3D ICs built using a foundry-grade 7nm bulk FinFET technology. Next, this study investigates the PPC tradeoff of 2-tier G-M3D ICs in the 7nm node, based

on the accurate wafer and die cost models along with the optimal CAD solution for two-tier G-M3D ICs. Each of these researches is organized into a self-contained chapter, and the key contributions of this dissertation are as follows:

Physical Designs for G-M3D ICs to Tackle FEOL/BEOL Degradation is presented in Chapter 2. Using a 7nm bulk FinFET from a foundry-grade process design kit (PDK), this chapter first models the mobility degradation of the top tier device caused by the low thermal budget process, and quantifies the impact of both W BEOL and cost-driven metal layer saving in the bottom tier on M3D design performance. Using these transistor corners and interconnect models, this work proposes Derated 2D flow, in which the geometry in technology files are not altered, but only the RC parasitics of interconnects are derated. Also a tier partitioning algorithm and post-partitioning optimization flow to tackle FEOL/BEOL degradation is introduced in Derated 2D flow. Experiments show that the proposed design solution allow only 3% performance degradation in G-M3D ICs compared to the timing result of 2D ICs under the worst FEOL/BEOL variation setting.

Power, Performance, Cost Tradeoff of G-M3D ICs is covered in Chapter 3. This chapter first develops highly-accurate full-chip, GDS-based wafer and die cost model for 2D and M3D ICs. Based on these cost modeling, this work optimizes the number of routing metal layers to obtain the best possible PPC values in 2D IC of two widely different circuit types (BEOL-dominant vs. FEOL-dominant). Next, this work proposes Projected 2D Flow that offers more than 50% footprint saving compared to that of 2D with the minimum design effort. Based on Projected 2D Flow, this work investigates PPC tradeoffs for two-tier, full-chip GDSII G-M3D ICs built using a foundry-grade 7nm bulk FinFET technology. Experiments reveal by how much cost should be further reduced to justify the adoption of M3D technology at the 7nm era.

Conclusion is discussed in Chapter 4. This chapter summarizes the researches presented in this dissertation and covers the future works that will improve the design quality of M3D ICs up to the commercial quality.

CHAPTER 2
DESIGN SOLUTION FOR G-M3D ICS TO TACKLE FEOL/BEOL DEGRADATION

Although an earlier work [12] addresses inter-tier performance variations in M3D ICs, this work is based on B-M3D ICs, in which the design seriously under-utilizes MIVs and is thus not practical. The authors of [11] present a design methodology for two-tier G-M3D ICs, so-called Shrunk 2D flow. The drawback of this flow is that the same RC parasitics are required despite shrinking geometry of layout objects. However, parasitics are changing non-linear along with metal geometry in the advanced technology nodes. Therefore, Shrunk 2D flow exaggerates the parasitic values, leading to low-quality M3D designs. Recently, M3D benefits have been studied for a predictive 7nm FinFET technology [15], but this work does not consider the inter-tier variation caused by limited thermal budget. In this chapter, physical design tools and methodologies for two-tier G-M3D ICs are developed to tackle the inter-tier performance variations caused by low temperature manufacturing. First, the top tier FEOL device mobility degradation and its impact on cell delay/power values are modeled. Next, the impacts of W interconnect and cost-driven metal layer saving in the BEOL of the bottom tier are quantified. Finally, these device and interconnect degradation models are used in the new full-chip G-M3D physical design flow named Derated 2D to show the power-performance savings of G-M3D ICs built using a foundry-grade 7nm FinFET technology under various FEOL/BEOL degradation settings.

Table 2.1 shows the nomenclature used in this research. We choose Triple Data Encryption Standard Cipher (DES3) and Low-Density Parity Check Decoder (LDPC) from OpenCore benchmark suites to cover different types of circuit. With regard to capacitance composition in Table 2.2, LDPC is a BEOL-dominant, and DES3 is a FEOL-dominant circuit. Figure 2.1 shows GDS layouts of their 2D implementation. All designs are imple-

Table 2.1: Nomenclature list used in this work.

TT	typical transistor corner (= no I_{on} degradation)
LT10p	10% I_{on} degradation in the top tier device
LT20p	20% I_{on} degradation in the top tier device
SVT	standard threshold voltage cell
LVT	low threshold voltage cell
Cu5	5 layers of copper BEOL used in the bottom tier
Cu3	3 layers of copper BEOL used in the bottom tier
W5	5 layers of tungsten BEOL used in the bottom tier
W3	3 layers of tungsten BEOL used in the bottom tier

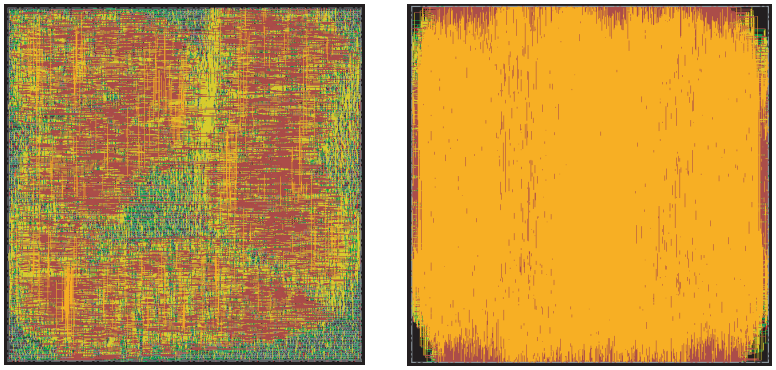

Figure 2.1: GDS layouts of 2D designs of the benchmarks.

mented with foundry-grade 7nm bulk FinFET process design kit (PDK). Using these two benchmarks, this work factorizes the impact of inter-tier variations caused by low temperature process on the performance of full-chip two-tier G-M3D designs. The diameter of an MIV is assumed to be the width of a top metal layer in the bottom tier (36nm for 5 metal BEOL, 24nm for 3 metal BEOL) with the resistance of 16Ω and capacitance of $0.01fF$.

2.1 FEOL/BEOL Variation Impact

2.1.1 Top Tier Device Degradation

The exact correlation between the low thermal budget process and the degree of top tier device degradation in the advanced node is not fully known. However, one of the main factors affected by low temperature process is expected to be the mobility of top tier devices. Therefore, in this work, the degree of top tier degradation are illustrated for two

Table 2.2: Benchmark circuits used in this work, where the metrics are from 2D IC designs. All designs are implemented with a foundry-grade 7nm bulk FinFET technology.

	DES3	LDPC
Cell Count	44,978	59,297
Wire Cap : Pin Cap	28:72	64:36
Avg Net Length (*um*/net)	2.54	10.02
Avg Net Wire Cap (fF/net)	0.41	1.77
Avg Net Pin Cap (fF/net)	1.03	0.97
Circuit Type	FEOL-dominant	BEOL-dominant

Table 2.3: Impact of mobility degradation on cell performance. Table shows the average output slew and delay in (ps) among INVx1, ND2x1, XNR2x1, AOI22x1, and DFF Clk-Q. Copper local interconnects are used.

	TT, Cu	LT10p, Cu	LT20p, Cu
Avg. SVT output slew	22.72 (1.00)	25.16 (1.11)	28.37 (1.25)
Avg. SVT cell delay	86.29 (1.00)	94.61 (1.10)	105.05 (1.22)
Avg. LVT output slew	16.62 (1.00)	18.27 (1.10)	20.32 (1.22)
Avg. LVT cell delay	57.28 (1.00)	62.90 (1.10)	69.81 (1.22)

scenarios, where 10% and 20% I_{on} decrease by mobility reduction. These degraded transistors are referred to LT10p, LT20p corner, respectively. In order to evaluate the impact of mobility degradation on a device, I_{on}, I_{off} of the device are measured by sweeping the mobility parameter. 7nm bulk FinFET Standard V_{TH} (SVT), and Low V_{TH} (LVT) compact models from a foundry-grade PDK are used with the nominal supply voltage 0.65V. From simulation results, it is observed that 18.2% and 33.4% of mobility degradation cause 10% and 20% I_{on} decrease, and 18.5% and 34.0% I_{off} reduction, respectively.

To analyze cell-level performance degradation, standard cell libraries are characterized for LT10p and LT20p corners with Cu local interconnect using Virtuoso Liberate. Using these cell models, output slew and gate delay of cells are measured assuming 10ps of the input slew and FO3 inverters with 300 Contacted Poly-Pitch (CPP $= 42nm$)-length M2 wireload using Synopsys PrimeTime. Table 2.3 shows that LT10p and LT20p corners result in 10.0%, 22.7% of cell performance degradation, respectively.

Starting from ideal scenario where there is no degradation on top tier devices and equivalent 5 metal layers of Cu BEOL in both tiers, Figure 2.2 shows the impact of top tier de-

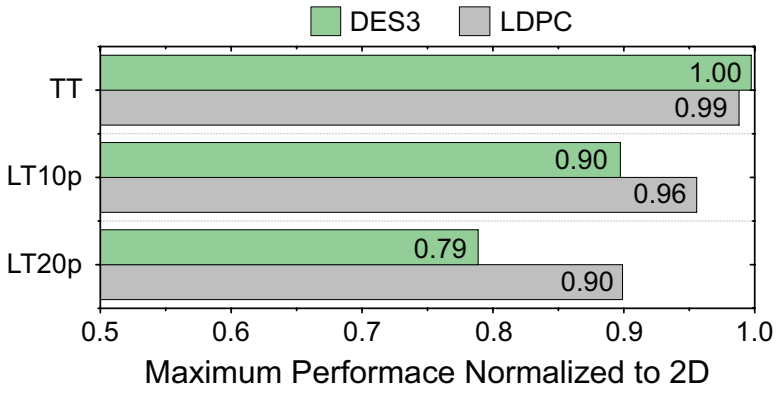

Figure 2.2: Impact of top tier device degradation on full-chip 2-tier M3D performance. 5 layers of Cu BEOL in both tiers are used. DES, the FEOL-dominant circuit, is more sensitive to the degradation.

vice degradation on the maximum performance of full-chip 2-tier gate-level M3D design. Shrunk 2D flow [11, 16] is used based on foundry-grade 7nm bulk FinFET PDK. Since Shrunk 2D flow does not handle the inter-tier variation, the last tier-by-tier routing stage is modified to consider the inter-tier performance variation as described in Section 2.2.3. Assuming 20% I_{on} reduction on the top tier, the performance of DES3, and LDPC is degraded by 21%, and 10% respectively. DES3 is a FEOL-dominant circuit, and 99% of the longest path delay consists of cell delays. Therefore, the performance of DES3 is sensitive to the FEOL degradation. On the other hand, LDPC has net delays of 21% out of the longest path delay, so the impact of cell delay increase is less than that of DES3. The simulation result identifies top tier device degradation as one of the critical obstacles to meet the timing of M3D design.

2.1.2 Bottom Tier Interconnect Degradation

In order to provide more thermal budget for the top tier manufacturing, integrating W BEOL in the bottom tier is an alternative. To quantify the impact of W interconnect, the process file (ICT) is modified from foundry-grade 7nm PDK assuming W conducting layers and TiN liners with the same geometry as Cu BEOL. Then the QRC technology file (TCH) is generated based on the process file using Cadence Techgen. Next, standard cell libraries

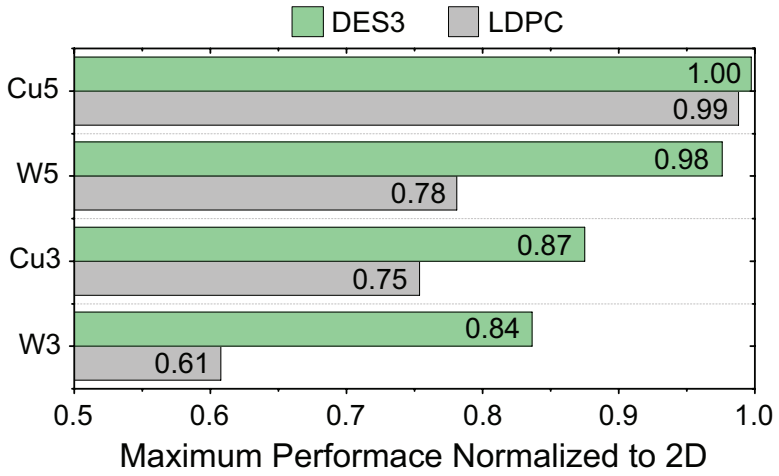

Figure 2.3: Full-chip impact of tungsten BEOL and metal layer saving in the bottom tier. LDPC, our BEOL-dominant circuit, is more sensitive to the changes.

are characterized based on W local interconnects using Virtuoso Liberate. Since wirelength of local interconnect is very short in the cell layout, only 2% slew degradation and 1% output delay increase in both SVT and LVT cells are observed. Based on these interconnect and cell models, Figure 2.3 shows the impact of tungsten interconnect in the bottom tier on the maximum performance of full-chip two-tier G-M3D designs. No device degradation on the top tier is assumed in this experiment. It is observed that the effective resistance of M2, M3 layers are 2.20 times as high as Cu resistance (ohm/um), and 2.46 times for that of M4, M5 layers. Comparing the maximum performance under 5 layers of Cu BEOL (Cu5) with the result under 5 layers of W BEOL (W5) case, LDPC has 21% performance degradation while the performance of DES3 is decreased by only 2%. This is because net delays of BEOL-dominant LDPC are significantly increased due to the highly resistive W interconnect. DES3 has minor performance degradation since most of the path timing consists of cell delays.

Another interesting perspective on the bottom tier BEOL is to reduce the number of metal layers. BEOL cost increases significantly from N28 to N7 nodes because of the multiple patterning processes [5]. Therefore, reducing metal stack must be taken into account to make M3D ICs cost-effective. 2 metal layers savings from the bottom tier are considered

in Figure 2.3. Comparing Cu3 result with Cu5 case, both DES3 and LDPC have significant performance loss. Reduced routing resources cause huge routing congestion in the bottom tier, resulting in 14% capacitance increase and 17% resistance increase on average. Under the worst scenario where W BEOL is used and 2 metal layers are reduced from the bottom tier, 16% and 39% of maximum performance is degraded in DES3 and LDPC, respectively.

2.2 Physical Design Solutions

To tackle the inter-tier performance variations caused by top tier low temperature manufacturing, this section proposes new full-chip M3D physical design flow named Derated 2D. Four CAD methodologies are proposed in Derated 2D flow as follows: (1) The layout objects are not modified but a 2D IC is designed with derated RC parasitic corner, named Derated 2D design. Then the placement result of Derated 2D design is projected into the final footprint of M3D design. (2) For the low-temperature process-aware tier partitioning, the cell slack is used as a metric for the timing criticality, and the timing critical elements are assigned into the bottom tier to address top tier cell degradation. (3) Timing-driven MIV planning deals with resistive W interconnect and reduced metal stack in the bottom tier. (4) A post-route optimization flow compensates the performance degradation under various FEOL/BEOL degradation settings at a minimum energy overhead. Overall design methodologies for Derated 2D flow is shown in Figure 2.4.

2.2.1 Derated 2D Design and Projection

Unlike Shrunk 2D flow [11] that requires shrinking of layout objects and RC parasitic scaling, Derated 2D uses original layout objects. However, Derated 2D is also possible to have overestimated wire load and redundant buffers, unless the wirelength savings from reduced footprint of M3D design is considered. Assuming no silicon area overhead, a two-tier M3D design has half footprint of that of 2D. In order to optimize the Derated 2D design with same RC parasitics of the M3D design, the RC corner is derated by $1/\sqrt{2}$ for total

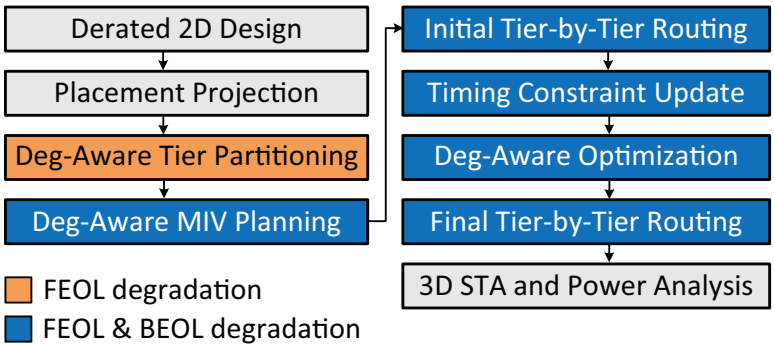

Figure 2.4: Derated 2D, a new FEOL/BEOL degradation-aware physical design flow for G-M3D ICs. The tier partitioning step tackles FEOL degradation, while the subsequent steps address both FEOL and BEOL degradation.

Table 2.4: Comparison between our Derated 2D flow and state-of-the-art Shrunk 2D flow [11].

	Derated 2D	Shrunk 2D
Shrink chip footprint?	No	Yes
Shrink cell layout?	No	Yes
Shrink metal dimension?	No	Yes
Scale unit-length RC parasitics?	Yes	Yes
Consider FEOL degradation?	Yes	No
Consider BEOL degradation?	Yes	No
Bottom tier cells use top tier metal?	Yes	No
Post-route optimization supported?	Yes	No

R and total C while not scaling coupling capacitance because of the same routing pitch in both Derated 2D and M3D. Then the whole placement result of Derated 2D design is projected into the footprint of final M3D design. Since every manhattan distance between each macro is scaled by $1/\sqrt{2}$ as a result of placement projection, RC parasitic of Derated 2D design is expected to be the same as that of M3D design. Table 2.4 and Figure 2.5 show comparison between Derated 2D flow and state-of-the-art Shrunk 2D flow.

2.2.2 Tier Partitioning and MIV Planning

Cell slack is a metric to measure how long each cell may delay without compromising the timing of paths propagating through the cell. Therefore, the slack value, which is a metric to represent the timing criticality of a cell, is extracted on the Derated 2D design

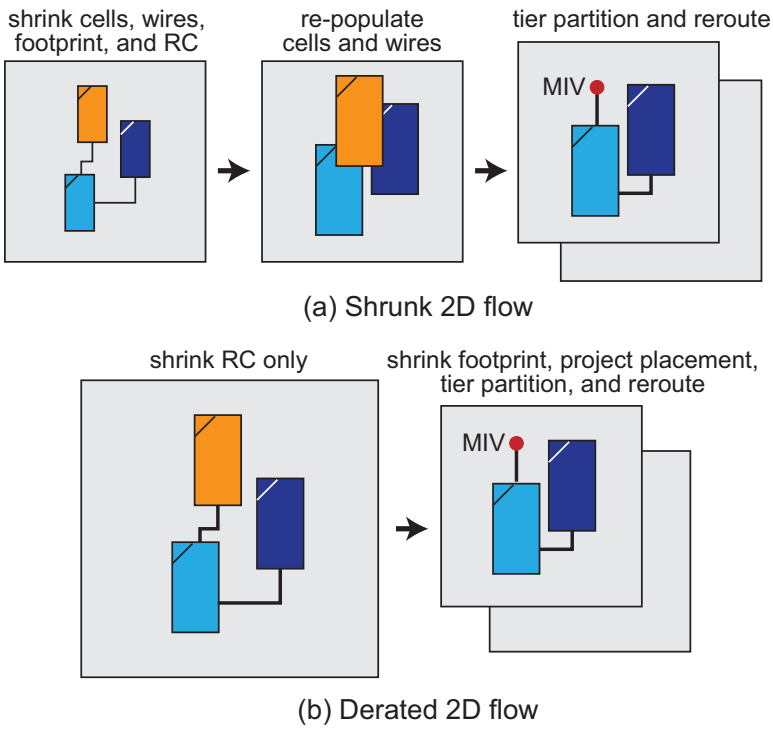

Figure 2.5: Illustration of Shrunk 2D [11] and Derated 2D flow.

using Synopsys PrimeTime. Clock network cells are kept in the bottom tier to avoid the change in the clock skew. The simplest partitioning scheme using cell slack is to sort them in decreasing order, and if the slack of a cell is less than median, then to place it on the bottom tier. In most cases, however, these timing critical cells are usually placed close to each other, resulting in local area skew between each tier. With unbalanced area skew, cell slack sorting does not guarantee minimum performance degradation since the original location of a cell that is already optimized at Derated 2D design must be changed during placement legalization. Therefore, the design footprint are divided by small size partitioning bin in the regular fashion, and cells within each bin are sorted by the slack value in decreasing order to meet the local area balance.

For the MIV planning, Shrunk 2D flow introduces CAD methodology to manipulate commercial engine built for 2D ICs [16]. The timing-driven MIV planning in Derated 2D flow also uses the basic idea of MIV planning scheme in Shrunk 2D flow but the differences are as follows: (1) In the same way that a 3D LEF is created to define two macro flavors

for each standard cell - one for each tier, a 3D LIB is created that defines two timing flavors to consider inter-tier device performance variation. Thus, timing model for each cell in 3D space is mapped to its appropriate transistor corner. The 3D LIB is imported into commercial router (Candence Innovus) to create delay corner for timing-driven routing. (2) Since each tier is possible to have different routing material and number of metal layers, a process file (ICT) and 3D TCH file are created for the full 3D metal stack. This 3D TCH file contains the RC parasitic information for every routing layers in M3D design. Then, a parasitic corner is created with this 3D TCH file in the commercial router. With timing constraint same as 2D design, timing-driven routing is done to insert MIVs. Using full 3D metal stack makes it possible to share the routing resources from all tiers. If the number of metal layers on the bottom tier is reduced, then the router uses top tier metal layers to route bottom tier nets in order to minimize routing congestion. If W BEOL is used in the bottom tier, the tool tries to use low resistive Cu BEOL on the top tier to minimize timing degradation. Figure 2.6 shows the differences of MIV planning scheme between Shrunk 2D and Derated 2D flow.

2.2.3 Post-Route Optimization and Routing

Since the initial Derated 2D design only involves normal transistor corner and Cu BEOL, it is clear that there exists limitation for timing closure under inter-tier variations in M3D design. Since delay and parasitic corners are created at timing-driven MIV planning stage, it is also possible to use a post-route optimization flow to update initial Derated 2D design for timing closure at a mimimum energy overhead. In order to do that, the size of macros in 3D LEF should be changed into the size of placement site, which is the smallest dimension that a macro can have. By using unit sized 3D macros, the placement overlap is removed or the cell legalization is minimized during post-route optimization. The location of initial top tier cells are the same, and the commercial tool is allowed to optimize bottom tier cells by resizing and VT swapping for timing closure. The reason why Derated 2D does not

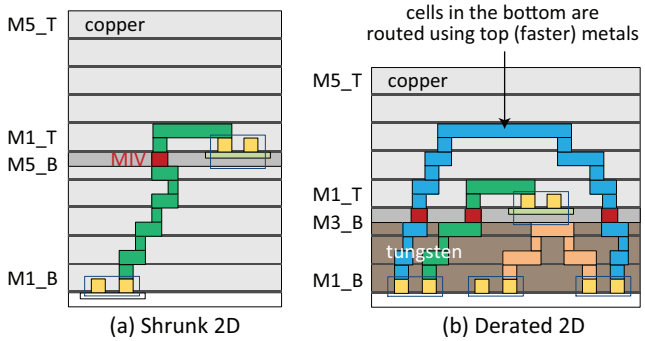

Figure 2.6: Metal stack comparison. (a) Shrunk 2D [11] with 5 Cu metal layers in both tiers, (b) Derated 2D flow with 5 layers of Cu in the top, and 3 tungsten in the bottom. Top cells contain MIV routing obstacle underneath.

play with the top tier cells is that the MIV routing blockages are initially fixed under the placement result of top tier cells. After post-route optimization, final tier-by-tier routing is proceeded to create separate GDS for each tier.

Once the MIV locations are determined by MIV planning, a DEF file is created for each tier containing the location of MIVs as primary I/O. Using original macro LEFs, the cell size is repopulated and the placement overlap is legalized. After the cells are routed initially with appropriate LIB (TT,LT10p,LT20p) and TCH (Cu,W) to the specific FEOL/BEOL degradation scenario, the timing context of each tier is created to optimize the routing quality. After routing under the timing context, the parasitic extraction and 3D timing and power analysis are followed.

2.3 Experimental Results

2.3.1 Impact of Tier Partitioning

Figure 2.7 shows the impact of cell-slack sorting tier partitioning on the design performance compared with Fiduccia-Mattheyes (FM) min-cut partitioning algorithm [11]. To be an equal comparison, Derated 2D designs are used for both partitioning algorithms, and 5 layers of Cu BEOL are assumed in both tiers. Even under 20% I_{ON} degradation on the top tier, cell-slack sorting partitioning allows only 5% of performance degradation in both

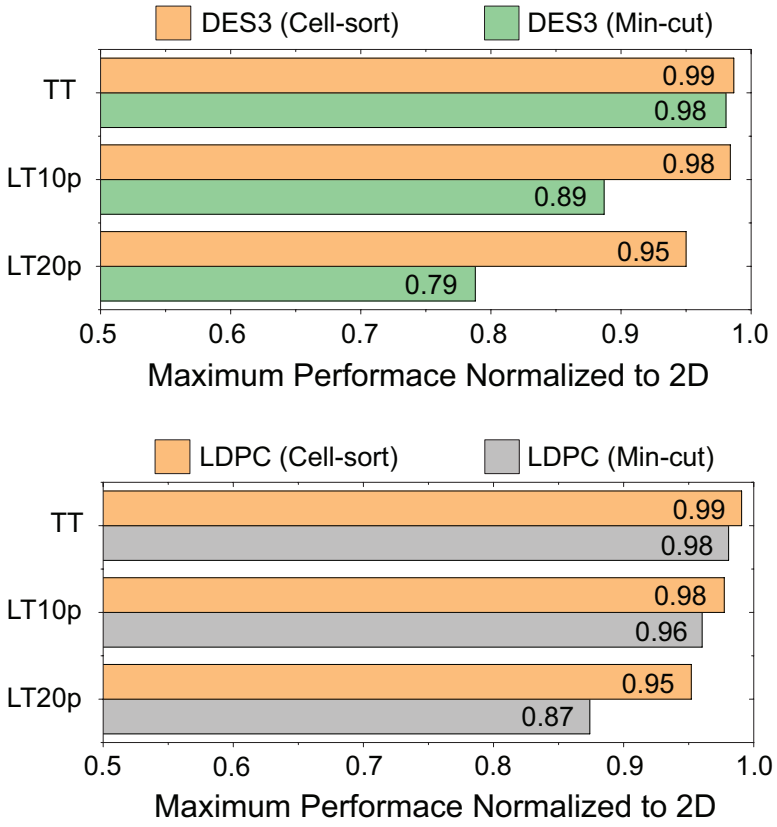

Figure 2.7: Tier partitioning impact on performance under FEOL degradation. The cell sorting-based method withstands the degradation better than min-cut for both circuits.

benchmarks. Table 2.5 shows detailed statistics of M3D designs from different partitioning algorithms. Min-cut partitioning tries to minimize the connections between each tier inside the partitioning bin. Therefore, 2D nets on each tier get longer and congested, leading to further longer 3D nets. However, cell slack sorting partitioning uses as many MIVs as necessary in order to assign the timing critical cells to the bottom tier. While minimizing the impact of top tier cell delay increase, these many and short 3D connections also effectively reduce net delay, resulting in significant timing saving. The incremental gain update makes FM min-cut heuristic run in $O(C)$, where C is the number of cells. Cell-slack sorting runs in $O(C \log C)$ by sorting algorithm.

Table 2.5: Comparison between cell-slack sorting vs. min-cut tier partitioning. LT20p transistor corner is used in the top tier, and 5 layers of Cu BEOL are used in both tiers.

	LDPC		DES3	
	min-cut	cell-sort	min-cut	cell-sort
Cell Count	57451		44805	
Net Count	59696		45036	
2D Net (top tier) Count	23171	16284	16677	10289
2D Net (bot tier) Count	24118	21718	23063	13701
3D Net Count	12407	21694	5296	21046
MIV Count	25958	37189	6772	25500
Avg. MIV# of 3D Net	2.09	1.71	1.28	1.21
Avg. WL of 2D Net (um/net)	3.29	2.83	2.18	1.69
Avg. R of 2D Net (ohm/net)	416.81	372.45	342.47	268.52
Avg. WL of 3D Net (um/net)	23.42	14.72	5.60	3.91
Avg. R of 3D Net (ohm/net)	2692.18	1743.30	867.53	650.95
Target Clock (ns)	1.0	1.0	0.5	0.5
WNS (s)	-0.16	-0.07	-0.18	-0.06
TNS (s)	-68.59	-2.86	-52.46	-3.58
TPS (s)	19.85	90.87	2605.71	2618.53
Runtime (sec)	24	111	12	44

2.3.2 Impact of MIV Planning

Based on cell slack sorting tier partitioning, Figure 2.8 shows the impact of the timing-driven MIV planning compared with Shrunk 2D flow. Under no top tier device degradation, W BEOL and 2 metal layer reduction in the bottom tier leads to 23% and 36% performance degradation in DES3 and LDPC with MIV planning in Shrunk 2D flow. The timing-driven MIV planing, however, allows only 13% and 20% of the performance degradation in DES3 and LDPC respectively. Table 2.6 analyzes net distribution and parasitics of LDPC design. Since 2D nets are possible to become 3D nets with the timing-driven MIV planning to close the timing, nets in the resistive bottom tier are routed by Cu BEOL in the top tier. Also, sharing routing resources between each tier decreases average net length and RC parasitics on the bottom tier and balances the routing congestion caused by metal layer reduction. Therefore, net delay degradation caused by W BEOL and routing congestion on the bottom tier are minimized, resulting in performance saving. In addition, top tier device degradation has a minor impact on design performance when bottom nets are in the worst

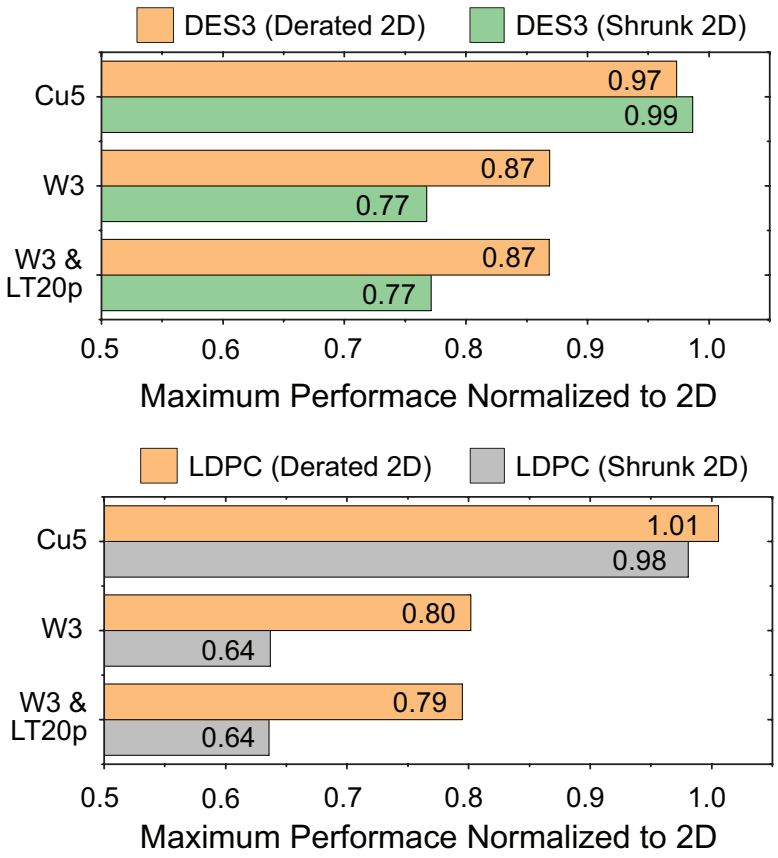

Figure 2.8: Impact of MIV planning in Derated 2D vs. Shrunk 2D [11]. Derated 2D withstands the FEOL and BEOL degradation better than Shrunk 2D.

scenario as a result of cell-slack sorting tier partitioning.

2.3.3 Comparison with the State-of-the-art

Based on a foundry-grade 7nm FinFET PDK, the design results of Derated 2D flow are compared with that of Shrunk 2D flow under all inter-tier variation scenarios in Table 2.7. Under the worst scenario, where there is 20% I_{on} reduction in the top tier while saving 2 metal layers of W BEOL in the bottom tier, Derated 2D result of LDPC achieves 36% of performance improvement, and 10% of energy saving compared with Shrunk 2D result without post-route optimization.

Comparing the design results between LDPC and DES3, it is observed that energy saving from M3D depends on the circuit type. LDPC, which is a BEOL-dominant circuit,

Table 2.6: Comparison of LDPC MIV planning result between Shrunk 2D [11] and Derated 2D, assuming no FEOL degradation and 3 tungsten BEOL layers in the bottom tier. Derated 2D encourages more routing in the top tier (= faster Cu BEOL).

	metric	Shrunk 2D	Derated 2D
net stats	top placed, top routed	17,432	17,410
	top placed, top/bot routed	0	22
	bot placed, bot routed	22,280	19,072
	bot placed, top/bot routed	0	3,208
	top/bot placed, top/bot routed	19,984	19,984
top tier	Avg. Net Length (um/net)	5.40	**6.85**
	Avg. Net Cap (ff/net)	2.70	**2.92**
	Avg. Net Wire Cap (ff/net)	0.92	**1.24**
	Avg. Net Res (Ohm/net)	601.81	**758.12**
bot tier	Avg. Net Length (um/net)	**3.50**	2.64
	Avg. Net Cap (ff/net)	**2.62**	2.45
	Avg. Net Wire Cap (ff/net)	**0.78**	0.52
	Avg. Net Res (Ohm/net)	**1192.32**	916.06
	Avg. MIV# per 3D net	2.1	1.6
	Max. Performance (GHz)	**0.68**	**0.75**
	Power-Delay Product (pJ)	32.59	32.22

has energy saving of 22% in ideal scenario, and still has 16% saving under the worst scenario without post-route optimization. This is because major source of energy saving from M3D design is wirelength reduction. In 2-tier M3D design with Derated 2D flow, expected maximum total wirelength saving is 29.3% considering 50% footprint saving from M3D design. Although the routing congestion in each tier is possible to degrade the wirelength saving depending on the circuit, this huge wirelength saving leads to around 30% of wire capacitance saving, and if the design is BEOL-dominant such as LDPC that 64% of total capacitance is wire capacitance, total capacitance saving becomes 22%. This capacitance saving is directly converted into switching power saving of 22%. However, since the total power consists of switching power, internal power, and leakage, the final power saving become 16%. In the case of DES3, wire capacitance is only 28% of total capacitance. Therefore, switching power saving from 30% of wirelength reduction in M3D is degraded into only 8%, and the final power saving degraded by internal power ratio become 2%.

The impact of a post-route optimization flow on timing and power-delay product is

also tabulated. Under any scenarios, the post-route optimization restores the performance degradation of M3D design up to minimum 97% of 2D performance. Under the worst scenario where 20% I_{on} degradation on the top tier and W BEOL with 2 metal layer saving in the bottom tier, Derated 2D recovers the M3D performance of LDPC from 79% to 97% of 2D performance at the expense of 8% of energy. In case of DES3, it is observed that although FEOL-dominant circuit has less energy saving, since it is less affected by resistive W interconnect and bottom routing congestion than BEOL-dominant circuit, it requires only 1% of 2D energy to restore the performance degradation.

2.4 Summary

This chapter proposed CAD methodologies for G-M3D ICs that tackle the FEOL/BEOL inter-tier variations caused by low temperature manufacturing. To address the top tier device degradation, a cell-slack sorting-based tier partitioning algorithm was presented to assign timing critical elements into the bottom tier. A timing-driven MIV planning flow and a post-route optimization flow were also developed to compensate the reduced routing layers and high resistance of tungsten interconnect. Experiments along with 7nm bulk FinFET from a foundry-grade PDK demonstrated that Derated 2D achieves up to 36% performance improvement and 10% energy saving compared with the state-of-the-art Shrunk 2D. Using a post-route optimization, Derated 2D further improved timing under the various FEOL/BEOL degradation settings at a minimum energy overhead.

Table 2.7: Performance and power-delay product (= energy) comparison under various FEOL and BEOL degradation settings. Our Derated 2D consistently outperforms Shrunk 2D [11] in terms of both performance and energy, even in the worst-case scenario (20% slow device, 3 layers of tungsten routing). Our post-route optimizer further improves performance at the expense of energy increase.

FEOL/BEOL setting		Maximum performance normalized to 2D			Post-route Optimization impact on TNS		Power-Delay Product normalized to 2D		
top tier	bot tier	Shrunk 2D	Derated 2D	D2D+PostOpt	Derated 2D	D2D+PostOpt	Shrunk 2D	Derated 2D	D2D+PostOpt
LDPC									
TT, Cu5	TT, Cu5	0.98	1.01	1.01	-0.01	-0.01	0.84	0.78	0.78
TT, Cu5	TT, Cu3	0.78	0.91	0.99	-13.06	-0.27	0.92	0.84	0.85
LT10p, Cu5	TT, Cu3	0.77	0.89	0.98	-16.05	-0.33	0.92	0.84	0.85
LT20p, Cu5	TT, Cu3	0.75	0.84	0.98	-38.12	-0.28	0.92	0.84	0.86
TT, Cu5	TT, W3	0.61	0.80	0.98	-137.90	-0.12	0.93	0.85	0.90
LT10p, Cu5	TT, W3	0.60	0.79	0.98	-159.11	-0.33	0.93	0.85	0.90
LT20p, Cu5	**TT, W3**	**0.58**	**0.79**	**0.97**	-186.12	-0.19	**0.93**	**0.84**	**0.92**
DES3									
TT, Cu5	TT, Cu5	1.00	0.97	1.03	-0.82	-1.08	0.98	0.98	0.98
TT, Cu5	TT, Cu3	0.87	0.90	1.02	-4.16	-1.78	1.01	0.99	1.00
LT10p, Cu5	TT, Cu3	0.86	0.90	1.03	-4.42	-1.43	1.01	0.99	1.00
LT20p, Cu5	TT, Cu3	0.79	0.90	1.03	-7.15	-2.32	1.01	0.99	1.01
TT, Cu5	TT, W3	0.84	0.87	1.01	-8.97	-2.77	1.02	1.00	1.01
LT10p, Cu5	TT, W3	0.82	0.87	1.03	-8.80	-2.44	1.02	1.00	1.01
LT20p, Cu5	**TT, W3**	**0.78**	**0.87**	**1.02**	-11.33	-1.96	**1.02**	**1.00**	**1.01**

CHAPTER 3

POWER, PERFORMANCE, AND COST TRADEOFF OF G-M3D ICS

Most of the earlier works on G-M3D ICs have focused on power, performance, and area improvement in the two-tier design given the same routing resources and silicon area as those of 2D ICs. For example, if a 2D IC has 5 metal layers and $100mm^2$ footprint, then a two-tier M3D IC has 5 metal layers and $50mm^2$ footprint on top and bottom tiers each. Based on those assumptions, [15, 8] shows that G-M3D ICs indeed offer huge iso-performance power savings compared with 2D ICs. Simply and ideally thinking, 50% footprint saving in M3D ICs results in 29.3% wire length reduction ($1/\sqrt{2}$ half perimeter wire length scaling) if the design aspect ratio is assumed to be the same [16]. This wire length savings not only decrease the wire capacitance (switching power savings) but also provides path timing margin to reduce buffer counts (internal power savings). Therefore, if the type of a design is a wire-dominant circuit, power savings in gate-level M3D are expected to be more.

However, since the footprint of wire-dominant circuits is determined by routability based on the limited routing resources, the design quality of this type of circuit would be easily improved when more routing layers are added. While M3D design needs to have the number of metal layers as few as possible to reduce the fabrication cost, adding more metal layers and optimizing BEOL metal stack in 2D IC can be easily achieved within a reasonable cost overhead [2]. Therefore, it leads to the next questions on how to set the proper 2D reference design for the fair PPC comparison with M3D design, and how much M3D has PPC-competitiveness to make us move toward the M3D era. This chapter addresses above questions.

3.1 Cost Modeling

Previous works [17, 18] on cost modeling for 3D IC are based on estimation of design parameters. Those studies use empirical constant for the area of standard cells, and expected wirelength distribution to predict total die area, and the number of required BEOL layers. In this research, accurate cost models are developed based on the real full-chip GDS design result.

Table 3.1: Nomenclature list used in this work.

$C_{W_{FEOL}}$	Manufacturing cost for FEOL		
C_{M_i}	Normalized manufacturing cost for metal layer M_i		
$C_{W_{BEOL,N}}$	Manufacturing cost for N BEOL layers		
$A_{W\|D}$	Wafer | Die area	$Y_{W\|D}$	Wafer | Die yield
D_W	Wafer defect density	DPW	# dies per wafer
$C_{W\|D_N}$	Wafer | Die cost for 2D IC with N BEOL layers		
$C_{W\|D_{N,M}}$	Wafer | Die cost for M3D IC with N (top) and M (bottom) BEOL layers		
α	Variable for M3D top tier manufacturing & bonding		
β	Variable for M3D wafer yield degradation		

3.1.1 Wafer Cost Model

Through the cost analysis framework from our industry partner, simple but self-contained wafer cost models are developed for 2D and M3D technology. Considering prescribed sequence of 7nm bulk FinFET process flow, and based on Cost-of-Ownership (CoO) where a database framework considers throughput of fab tools, material, labor, repair, utility and overhead expenses due to the equipment operation [4, 5], the ratio between FEOL and BEOL manufacturing cost is set as 30%:70%. 2D BEOL metal stack configuration used in this paper is in accordance with International Technology Roadmap for Semiconductor (ITRS) guidelines for 7nm technology node. Since the foundry-grade 7nm bulk FinFET device technology is assumed to have the middle of line (MOL), MINT layer is included in the metal stack, but it is only used for intra cell routing.

Table 3.2: Assumed patterning option and manufacturing cost per metal layer.

Layer	Patterning	Pitch	Width	Thickness	Normalized Cost (C_{M_i})
MINT (M0)	SAQP	32nm	21nm	24nm	2.8
M1	LELE	42nm	24nm	24nm	1.7
Mx	SAQP	32nm	24nm	24nm	2.8
My	LELE	48nm	24nm	48nm	1.5
Mz	LE	80nm	40nm	80nm	1.0

Table 3.2 shows the assumed patterning option and manufacturing cost per metal layer (C_{M_i}) obtained from industry partner. Manufacturing costs for MEOL and intermediate interconnect layers are normalized with the cost for global interconnect layer (Mz). With this Table and proposed ratio between FEOL and BEOL manufacturing cost, the reference design is set as 2D IC with 8 BEOL metal layers, and the normalized wafer cost is calculated for another designs with different metal stack as shown below.

$$C_{W_{FEOL}} = 0.3 \times C_{W_8}, C_{W_{BEOL,8}} = 0.7 \times C_{W_8}$$

$$C_{W_{BEOL,N}} / C_{w_{BEOL,8}} = \sum_{i=0}^{i=N} C_{M_i} / \sum_{i=0}^{i=8} C_{M_i}$$

$$C_{W_N} / C_{W_8} = (C_{W_{FEOL}} + C_{W_{BEOL,N}}) / C_{W_8}$$

2D Wafer Cost Model: For N BEOL metal layers,

$$C_{W_N} / C_{W_8} = 0.3 + 0.7 \times \sum_{i=0}^{i=N} C_{M_i} / \sum_{i=0}^{i=8} C_{M_i} \tag{3.1}$$

In literature, no work has previously studied cost estimation for M3D integration. Cost for sequential integration is not fully known yet, and top tier manufacturing should be limited due to the FEOL and BEOL integrity on the bottom tier. Therefore, in this work, it is assumed that the FEOL cost for both tiers are the same as default, and a variable is included to take into account the different device manufacturing cost in each tier and bonding cost (α). M3D BEOL cost is calculated by the sum of BEOL cost for each tier.

M3D Wafer Cost Model: For N (top) and M (bottom) BEOL metal layers,

$$C_{W_{N,M}}/C_{W_8} = 0.6 + \alpha + 0.7 \times (\sum_{i=0}^{i=N} C_{M_i} + \sum_{i=0}^{i=M} C_{M_i}) / \sum_{i=0}^{i=8} C_{M_i} \tag{3.2}$$

3.1.2 Die Cost Model

Considerations for the cost of I/O pins, packaging, testing, and cooling are out of the scope in this paper. Assuming that edge clearance and notch height of the wafer are ignorable, the die manufacturing cost takes into account the number of dies per wafer, die yield, and die area. For M3D die yield, sensitivity variable β are multiplied to 2D wafer yield, so that it leads to evaluating how much M3D wafer yield should be improved to guarantee the M3D benefits compared with 2D. Experiments are done with $300mm$ of wafer diameter, and $0.2mm^{-2}$ of D_W, and 0.95 of Y_W. Finally,

2D Die Cost Model: For N BEOL metal layers,

$$DPW_N = A_W/A_{D_N} - \sqrt{2\pi A_W/A_{D_N}} \tag{3.3}$$

$$Y_{D_N} = Y_W \times (1 + A_{D_N} D_W/2)^{-2} \tag{3.4}$$

$$C_{D_N}/C_{D_8} = \frac{C_{W_N}}{C_{W_8}} \times \left(\frac{DPW_8 \times Y_{D_8}}{DPW_N \times Y_{D_N}} \right) \tag{3.5}$$

M3D Die Cost Model: For N (top) and M (bottom) BEOL metal layers,

$$DPW_{N,M} = A_W/A_{D_{N,M}} - \sqrt{2\pi A_W/A_{D_{N,M}}} \tag{3.6}$$

$$Y_{D_{N,M}} = \beta \times Y_W \times (1 + A_{D_{N,M}} D_W/2)^{-2} \tag{3.7}$$

$$C_{D_{N,M}}/C_{D_8} = \frac{C_{W_{N,M}}}{C_{W_8}} \times \left(\frac{DPW_8 \times Y_{D_8}}{DPW_{N,M} \times Y_{D_{N,M}}} \right) \tag{3.8}$$

3.2 Physical Design Solutions

In [11], authors present state-of-the-art Shrunk 2D flow for full-chip GDS gate-level monolithic 3D IC. The idea of this design flow is to manipulate the powerful optimization capability of the commercial tool built for 2D ICs at pseudo-3D design environment where shrunk layout objects are placed and routed in the floorplan with the same dimension as final M3D. For example, assuming 2-tier, gate-level M3D design with zero silicon area overhead, the footprint of each tier should become 50% of 2D design footprint. For the Shrunk 2D flow, first the floorplan size is fixed as same as the footprint of final M3D, and shrink the geometric dimension of original 2D layout objects to scale by $\sqrt{2}$. Then the area of standard cells become 50% of original cell area, and also the pitch and width of interconnects become 70.7% of the original. Now, the unit-length RC parasitic is also scaled to let commercial router use the original parasitic of interconnects in optimization stages. This scaling procedure is necessary to remove the overlap between standard cells in the shrunk chip footprint, and to obtain reasonable timing optimization by commercial tool.

However, shrinking layout objects is subject to Design-Rule-Violations (DRV) in the complicated standard cell layouts in the advanced technology nodes. Also, scaling RC parasitics of shrunk interconnects to match the parasitics same as the original either is incorrect due to the exaggerated extrapolation of parasitics with internal algorithm in the commercial tool, or requires many efforts to modify the geometric and electrical characteristics in the interconnect files. Furthermore, those layout objects are not reusable in the design with more than 2-tiers. Lastly, Shrunk 2D Flow possibly maximizes the placement utilization of each tier in M3D design, but it does not fully optimize the design in terms of routing utilization since reduced footprint and effectively routed nets in M3D decreases total wirelength. Therefore, a new physical design solution named Projected 2D is proposed for two-tier gate-level M3D designs. The main idea of this flow is to use 2D design itself as a starting point for implementation of M3D design. The overall design steps are shown in

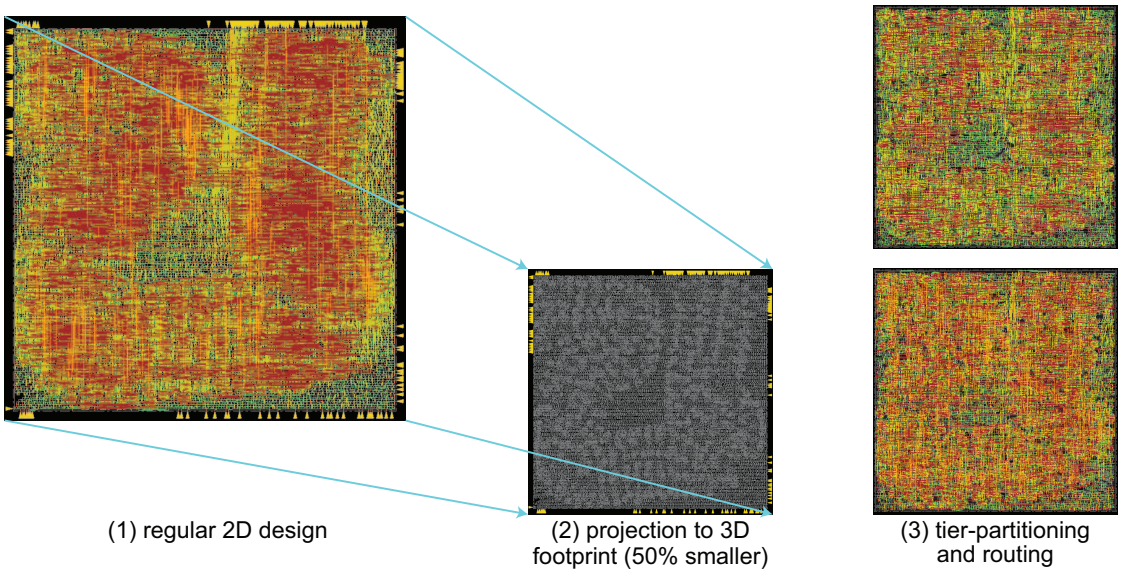

Figure 3.1: Major steps of our Projected 2D flow. (a) regular 2D IC design, (b) placement projection, (c) tier partitioning and tier-by-tier routing after MIVplanning.

Figure 3.1.

3.2.1 Projected 2D Flow

Projected 2D does not require shrinking of layout objects, and scaling RC parasitics unlike Shrunk 2D flow. The beauty of Projected 2D flow is as follows: (1) After 2D design is implemented, which already closes design specification with normal process-design-kit (PDK), Projected 2D uses final netlist and placement result of 2D design to implement M3D design. Since there is no difference between the netlist of 2D and that of M3D, it is possible to directly compare the routing result of equivalent nets. Analyzing RC parasitics of the nets through comparison with 2D, it allows us to confirm the wirelength saving from M3D, or to improve tier partitioning result for better M3D design quality. (2) Projected 2D maximizes either placement or routing utilization by projection of 2D placement result. Modulating the projection factor, the final M3D design footprint can be easily reduced by more than 50% if there is enough routing usage saving. (3) Projected 2D enables multi-tier, gate-level M3D design without any efforts for modification of geometric information in input design files.

Table 3.3: Comparison between Projected 2D and state-of-the-art Shrunk 2D flow [11].

	Projected 2D	Shrunk 2D
Shrink macro layout?	No	Yes
Shrink interconnect dimension?	No	Yes
Scale unit-length RC parasitics?	No	Yes
Consider buffer saving in M3D?	No	Yes
Have same netlist as 2D?	Yes	No
Maximize routing utilization?	Yes	No
LDPC M3D result, 7nm bulk FinFET, M5 (top) / M5 (bottom)		
Chip Area (μm^2)	4499	5408
Maximum routing utilization	0.762	0.666
Total buffer count	16163	15980
Total power (mW)	32.76	32.41
WNS (ns)	0.057	-0.015

However, Projected 2D overestimates wire loads, and inserts redundant buffers during 2D optimization by commercial tool. Table 3.3 shows qualitative, and quantitative comparison between Projected 2D and Shrunk 2D. Assuming 2-tier LDPC M3D design with a foundry-grade 7nm bulk FinFET PDK and 5 metal layers in both tiers, design result of Projected 2D flow has more buffers resulting in larger positive slack than that of Shrunk 2D. On the other hand, due to the reduced footprint of Projected 2D design, it has more wirelength saving and consequent switching power saving to compromise increase in internal power due to redundant buffer counts.

3.2.2 Tier Partitioning and MIV planning

Based on projected placement location of macros and netlist, placement-driven min-cut partitioning is used for the tier partitioning [16]. This partitioning scheme divides the whole design in regular fashion for the balanced local area skew, so-called partitioning bin, and do Fiduccia Mattheyses (FM) min-cut partitioning inside each of partitioning bins. Therefore, the number of inter-tier connections depends on the size of partitioning bins. In [8], it is shown that there is an optimization point for the minimum power consumption along with the inter-tier connections. This is because too many 3D connections cause routing congestion and redundant snaking between each tiers, while few 3D connections leads to

small wirelength savings. Therefore, the best partitioning bin size per benchmark is found by sweeping the size of partitioning bin for the maximum power savings.

After tier partitioning, the proper MIV location is located by using commercial tool built for 2D ICs as proposed in [16]. The main idea of this methodology is to let commercial router treat MIVs same as normal vias while there are routing blockages on the area of macros on the top tier to avoid overlap between MIVs and top macros during routing stage. The limitation of this flow is that the direction of metal layers should not be the same between adjacent layers, and that the number of interconnect layers on the bottom tier should be an even number. Since our foundry-grade 7nm bulk FinFET standard cell layout contains MINT layer for internal routing, an odd number of interconnect layers on the bottom tier is assumed.

Once the MIV locations are determined by MIV planning, a DEF file for each tier is created containing the location of MIVs as primary I/O. Then the timing context of each tier is created to optimize the routing quality. After routing under the timing context, RC parasitics are extracted, and 3D timing and power analysis is proceeded by using Synopsys Primetime.

3.2.3 Footprint Resizing

Once initial M3D design is done, the maximum placement or routing utilization is checked on each tier if it is over 70%. Since M3D placement utilization on each tier is same as 2D placement utilization considering balanced area skew from placement-driven min-cut partitioning, meeting the sufficient placement utilization is guaranteed from 2D design result. However, if a circuit is BEOL-dominant type, then 2D placement utilization is possible to be lower than 70% because insufficient routing resources requires large die area. In that case, even though 2D routing utilization is over 70% in certain metal layers, routing utilization in M3D could be lower than 70% due to the wirelength reduction. To maximize the utilization of die area, the proper footprint is estimated as $A'_D = A_D \times U_r/0.7$, where

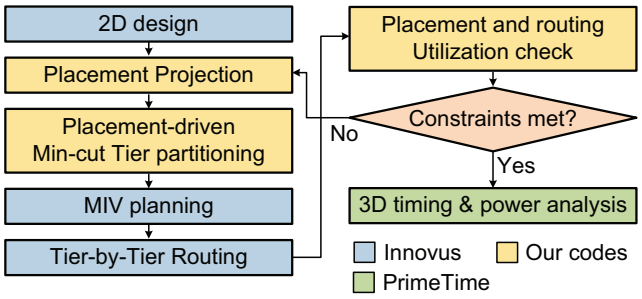

Figure 3.2: Projected 2D design flow.

U_r is the maximum routing utilization out of all metal layers, A_D is the current footprint area, and A'_D is the updated footprint area. We project the 2D placement into the updated footprint, and iterate the design flow shown in Figure 3.2 until the U_r is over 70%.

3.3 Experimental Results

We choose Triple-Data-Encryption-Standard cipher (DES3) and Low-Density Parity-Check decoder (LDPC) from OpenCore benchmark suites to cover two widely different circuit types. 2D Design of these two benchmarks built using a foundry-grade 7nm bulk FinFET PDK shows that 72% of total capacitance in DES3 is pin capacitance while 64% of total capacitance in LDPC is wire capacitance. Also, average net length of LDPC is 3.94 times longer than that of DES3. Therefore, LDPC is defined as a BEOL-dominant circuit, and DES3 as a FEOL-dominant circuit. The diameter and pitch of an MIV in the experiments is assumed to be 24nm and 48nm with resistance of 16Ω and capacitance of $0.01fF$.

Table 3.4: 2D IC PPC analysis and comparisons. Our PPC is defined in Equation 3.9.

Circuit Type	Metal Stack	Tot. Power (mW)	Max. Perf (GHz)	Placement Utilization	Max. Routing Utilization	Wafer Cost	Area (μm^2)	DPW (1e+6)	Die Yield	Die Cost	PPC
FEOL dominant DES3	M3	37.70	2.00	0.719	M2, 0.287	0.739	6048	11.679	0.949	0.739	**1.306 (best)**
	M4	36.96	2.00	0.718	M3, 0.242	0.804				0.804	1.224
	M5	36.52	1.99	0.716	M3, 0.215	0.870				0.870	1.140
	M6	36.69	2.00	0.716	M3, 0.213	0.913				0.913	1.086
	M7	36.39	1.99	0.716	M3, 0.214	0.957				0.957	1.040
	M8	36.21	1.99	0.715	M3, 0.207	1.000				1.000	1.000
BEOL dominant LDPC	M5	39.28	0.99	0.359	M4, 0.824	0.870	10816	6.529	0.948	1.720	0.433
	M6	33.45	0.99	0.581	M6, 0.807	0.913	6561	10.765	0.949	1.094	0.799
	M7	31.49	0.99	0.686	M6, 0.790	0.957	5476	12.899	0.949	0.957	0.972
	M8	29.28	0.99	0.794	M8, 0.613	1.000	5476	12.899	0.949	1.000	1.000
	M9	28.39	0.99	0.787	M8, 0.678	1.043	4692	15.055	0.949	0.894	1.154
	M10	27.48	1.00	0.789	M4, 0.535	1.087	4692	15.055	0.949	0.931	**1.156 (best)**

3.3.1 2D Design Results

Table 3.4 shows the impact of changing metal stack configuration on the design result of FEOL-dominant circuit DES3, and BEOL-dominant circuit LDPC. Designs for each benchmark are constrained with the same clock period, (0.5ns for DES3, 1.0ns for LDPC). Total power in the Table 3.4 is iso-performance power number, and the maximum performance is calculated by reversing the sum of clock period and the worst timing slack. PPC is calculated as follows:

$$PPC = \frac{Max\_Performance}{Total\_Power \times Die\_Cost} \tag{3.9}$$

Since wafer and die cost is normalized with that of 8 BEOL metal stack (M8 in Table 3.4) design, PPC is also normalized with the PPC value of M8 design.

FEOL-Dominant Circuit Type

Starting from M8 design, reducing metal layers in FEOL-dominant circuit has little impact on routing utilization overhead. Since most of nets in DES3 is locally routed, maximum routing utilization is only 20.7% in M3 layer even though there are 8 BEOL metal layers for routing. The placement utilization and die area are also unchanged along with metal stack reduction since M3 design already has sufficient routing resources. All designs close the timing, and small change in iso-performance power along with metal stack reduction is caused by slightly increased routing congestion. Even though the total power in M3 design is increased by 4% compared to M8 design, wafer and die costs are reduced by 26%. Therefore, overall PPC saving of M3 design is 31% more than the saving of M8 design, and M3 design is defined as the most optimized design for DES3 in terms of PPC.

BEOL-Dominant Circuit Type

BEOL-dominant circuit LDPC shows interesting results in Table 3.4. In M5 design, the die area is determined by the maximum routing utilization in M4 layer. The lack of routing resources increase chip size even though placement utilization is only 35.9%. The large footprint not only increases die cost, but also makes overall wirelength longer and leads to higher wire capacitance. Therefore, adding only one more metal layer significantly improves the design quality of BEOL-dominant circuit. Compared to M5 design result, M6 design has total power saving by 15%, area reduction by 39%, lower die cost by 36%, and PPC improvement by 85%.

Once there are enough metal layers for the routing in LDPC, the die area needs to be determined by both placement and routing utilization. Therefore, area saving and the impact of adding more interconnect layers become saturated as shown in M8 design. As a result, reduced power saving and additional cost for more metal layer have a tradeoff relationship.

3.3.2 Impact of Metal Stack Optimization

Optimizing dielectric constant, and conductivity in the metal stack by changing material composition is one of the cheapest solutions to improve design quality. We assume that the dielectric constant of global interconnect layers (from M6 to M10) has been reduced by 14%, and generate new technology file (TCH) using Cadence Techgen. Scaling dielectric constant reduces 12% of total capacitance per unit length for the global interconnect metal layers, and this metal stack configuration is defined as Low-K metal stack. We also consider the wafer cost change for the Low-K metal stack. Based on the wafer cost model in Section 3.1, the BEOL cost is increased from 0.70 to 0.71 and takes it into account for the PPC calculation.

Table 3.5 shows the impact of Low-K metal stack on the BEOL-dominant LDPC 2D designs. By comparing M5 design with M5 + Low-K design, reduced wire capacitance by

Table 3.5: Impact of Low-K metal stack on BEOL-dominant LDPC 2D designs.

Metal Stack	Tot. Power (mW)	Max. Perf (GHz)	Wafer Cost	Area (μm^2)	Die Cost	PPC
M5	39.28	0.99	0.870	10816	1.720	0.433
M5 + Low-K	37.27	0.99	0.878	8190	1.314	0.598
M6	33.45	0.99	0.913	6561	1.094	0.799
M6 + Low-K	32.4	0.99	0.922	6561	1.105	0.818
M7	31.49	0.99	0.957	5476	0.957	0.972
M7 + Low-K	30.72	0.99	0.966	5476	0.966	0.987
M8	29.28	0.99	1.000	5476	1.000	1.000
M8 + Low-K	28.35	0.99	1.010	4692	0.865	**1.194**
M9	28.39	0.99	1.043	4692	0.894	1.154
M9 + Low-K	27.56	1.00	1.054	4692	0.903	1.188
M10	27.48	1.00	1.087	4692	0.931	1.156

using Low-K metal stack further improves total power due to the switching power saving. Also, decreased routing congestion from the reduced number and drive strength of buffers make room for die area saving. Since it is assumed that BEOL cost for Low-K metal stack is different from the normal metal stack, it shows different tradeoff between power saving and wafer cost increase. Even though M9 + Low-K design has more power saving than M8 + Low-K design, the PPC value of M8 + Low-K is higher than M9 + Low-K due to the BEOL cost. Overall, M8 + Low-K design is defined as the most optimized design for LDPC with regard to PPC.

For the FEOL-dominant DES3 design, the impact of reducing wire capacitance on PPC by using Low-K metal stack is negative since it has little power saving with increased die cost.

3.3.3 M3D Design Results

Table 3.6 shows the M3D design results using normal metal stack of various combinations. 2D design in Table 3.6 is the best design with regard to PPC, defined as the reference for the comparison with M3D design. In this section, the variable for the sequential integration and bonding cost for the top tier (α) is assumed as 0.1, and M3D wafer yield (β) as 90% of 2D wafer yield.

Table 3.6: M3D PPC analysis and comparison. Our PPC is defined in Equation 3.9. Power is total power consumption, and Perf is the maximum performance.

Circuit Type	Design Flavor	Metal Stack (top / bottom)	Power (mW)	Perf (GHz)	Placement Utilization (top / bottom)	Max. Routing Utilization (top / bottom)	Wafer Cost	Area (μm^2)	DPW (1e+6)	Die Yield	Die Cost	PPC
FEOL dominant DES3	2D	M3	37.7	2	0.719	M2, 0.287	0.739	6048	11.679	0.949	0.739	1.306
	M3D	M3 / M5	37.38	1.776	0.744 / 0.718	M2, 0.278 / M4, 0.215	1.826	3041	23.232	0.854	1.019	0.848
		M4 / M5	36.83	1.901	0.745 / 0.718	M3, 0.203 / M4, 0.215	1.872			0.854	1.045	**0.899**
		M5 / M5	36.74	1.901	0.745 / 0.718	M3, 0.187 / M4, 0.215	1.917			0.854	1.070	0.880
		M6 / M5	36.74	1.898	0.745 / 0.718	M3, 0.187 / M4, 0.215	1.948			0.854	1.087	0.864
BEOL dominant LDPC	2D	M8 + Low-K	28.35	0.99	0.794	M8 0.713	1.010	4692	15.055	0.949	0.865	1.194
	M3D	M5 / M5	32.76	1.060	0.481 / 0.425	M4, 0.762 / M4, 0.639	1.917	4499	15.702	0.854	1.750	0.547
		M6 / M5	32.55	1.050	0.563 / 0.491	M6, 0.666 / M4, 0.679	1.948	3894	18.142	0.854	1.538	0.620
		M7 / M5	32.37	1.018	0.563 / 0.491	M6, 0.631 / M4, 0.694	1.978	3894	18.142	0.854	1.562	0.596
		M5 / M7	28.5	1.035	0.606 / 0.528	M4, 0.756 / M4, 0.545	1.978	3504	20.162	0.854	1.406	**0.764**

Table 3.7: Equivalent net comparison between M3D and 2D design. The worst resistance net in DES3 M3D design is analyzed.

Wirelength distribution (um)	2D	M3D (top/bottom)
M5	122.35	0.00 / 74.90
M4	67.09	7.42 / 51.74
M3	0.32	5.87 / 3.78
M2	0.27	2.46 / 0.19
M1	0.46	1.50 / 0.46
Net Total Wirelength (μm)	190.50	148.05
Net Total Resistance (Ω)	11187	10206
Unit-length Resistance ($\Omega/\mu m$)	58.72	68.94

FEOL-Dominant Circuit Type

While 2D DES3 design with only M3 metal stack already has enough resources to finish the routing, M3D DES3 design should have M5 metal stack in the bottom tier. This is because if M3 metal stack is used in the bottom tier, part of routing resource in M3 layer will be dedicated to inter-tier connection (MIV planning), resulting in compromise of routability, and many DRVs. Also, due to the limitation of MIV planning scheme using commercial 2D router, the odd number of BEOL metal layers is allowed on the bottom tier so that top metal layer of the bottom tier and MINT layer of the top tier has routing direction orthogonal to each other. Therefore, 5 metal layers are set as the minimum metal stack on the bottom tier, and evaluate the PPC benefit of M3D design.

Since the die area of FEOL-dominant circuit is determined by placement utilization, 2-tier M3D DES3 design indeed has 50% of footprint saving compared to 2D design. However, high wafer cost of M3D integration, and the assumptions on reduced M3D wafer yield increase die cost for M3D. In addition, total power saving in M3D is not significantly large, since DES3 is FEOL-dominant and most of routing in DES3 are done locally. Performance loss in DES3 M3D design is worth to notice. Because M3D design keeps the same nets as 2D design through Projected 2D flow, the worst resistance net in M3D design is compared with the equivalent net in 2D design as shown in Table 3.7.

It shows detailed wirelength distribution and net resistance of the equivalent net in 2D

M5 design and M3D M5 / M5 design. The 2D net has long wirelength, but most of routing are done in M5 layer. However, although the M3D net has 22% total wirelength saving, total net resistance is reduced by 9% only. Unit-length resistance of the M3D net is 17% higher than that of the 2D net. Based on the net comparison, it is observed that when locally placed and routed cells in 2D design are split into different dies through tier partitioning, routing utilizations for intermediate interconnect layers are increased. Since part of the top metal layer in the bottom tier should be dedicated to MIV planning, commercial router is not able to fully use the top metal routing resource in the bottom tier. Instead, it uses more intermediate interconnect layers. Besides, wires should go through the whole metal stack in the bottom tier to route top tier cells. Therefore, it is likely to increase the routing congestion, and redundant wire capacitance.

Furthermore, top tier routing also uses intermediate interconnect layers since only local routing remains. The resistance of M2,M3 layer is 2.46 times higher that of M4,M5 layer. Therefore, locally routed FEOL-dominant circuit requires more effective tier partitioning, otherwise the timing of the critical path worsens. With regard to PPC value, M3D design with M4 / M5 metal stack is defined as the most optimized M3D design for FEOL-dominant DES3.

BEOL-Dominant Circuit Type

When comparing M3D M5 / M5 design with 2D M5 design, BEOL-dominant LDPC M3D design indeed has increases of power savings by 17% and die area savings by 58%. However, in Section 3.3.2, 2D M8 + Low-K design is defined as the reference design for the fair comparison with M3D designs. Since placement and routing utilization of our 2D reference design is highly optimized, die area of 2D design is small enough to provide cheap die cost. Due to the die area as small as 57% of 2D M5 design, huge wirelength and buffer saving result in M3D-compatible power consumption.

Therefore, unlike FEOL-dominant DES3 M3D design, LDPC M3D M5 / M7 design

Table 3.8: Impact of Low-K metal stack on BEOL-dominant LDPC M3D designs.

Metal Stack (top / bottom)	Tot. Power (mW)	Max. Perf (GHz)	Wafer Cost	Area (um)	Die Cost	PPC
2D						
M8 + Low-K	28.35	0.99	1.010	4692	0.865	1.194
M3D						
M5 / M5	32.76	1.060	1.917	4499	1.750	0.547
M5 / M5 + Low-K	32.12	1.074	1.929	4499	1.760	0.562
M6 / M5	32.55	1.050	1.948	3894	1.538	0.620
M6 / M5 + Low-K	31.9	1.071	1.960	3894	1.548	0.641
M7 / M5	32.37	1.018	1.978	3894	1.562	0.596
M7 / M5 + Low-K	31.72	1.031	1.991	3894	1.572	0.611
M5 / M7	28.5	1.035	1.978	3504	1.406	0.764
M5 / M7 + Low-K	27.91	1.050	1.991	3504	1.414	**0.787**

is the best M3D design out of given metal stack combinations with regard to PPC value though it only has 25% area saving compared with 2D reference. Table 3.8 shows the impact of Low-K metal stack on LDPC M3D design. By using Low-K metal stack, M5 / M7 + Low-K design finally beats 2D reference in terms of both total power and maximum performance. Even though it is clear that using Low-K metal stack and adding routing resources are very effective solutions to improve M3D design quality, too much expensive metal stack for BEOL-dominant circuit increases the wafer cost almost 2 times higher than 2D reference, resulting in lower PPC of M3D than that of 2D.

3.4 7nm M3D Cost and Yield Study

In Section 3.3.3 and 3.3.3, assuming M3D wafer yield (β) as 90% of 2D wafer yield, and additional cost for top tier device implementation (α) is 10% of wafer cost for 2D M8 design, it is observed that PPC of FEOL-dominant DES3 M3D design is worse by 31% and BEOL-dominant LDPC M3D design by 34% compared to 2D reference. Then the next question is how much M3D wafer yield and additional cost for M3D integration should be further reduced for the cheap M3D die cost to justify the adoption of M3D technology. In Figure 3.3, red surface of each plot shows the valid region along with α, and β where the best M3D design defined in the previous Sections beats PPC of the 2D reference. Z-axis

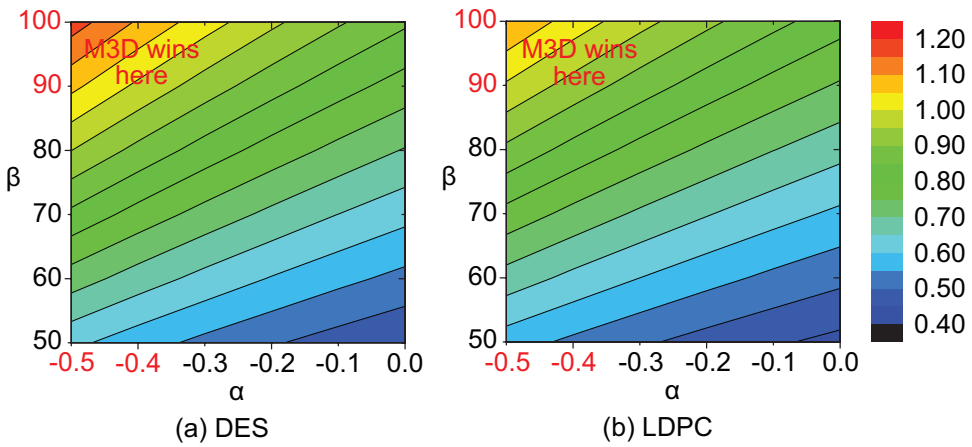

Figure 3.3: M3D cost vs. yield vs. PPC sensitivity analysis. α denotes cost variable for top-tier devices fabriacation and bonding in M3D, e.g., α = -0.4 means that FEOL manufacturing cost for M3D (0.6) should be 67% lower (0.6 + α = 0.2). β denotes M3D wafer yield (percentage w.r.t. 2D wafer yield). Z-axis denotes PPC ratio of M3D over 2D, e.g., 1.2 means M3D PPC is 20% better.

of these plots is calculated by the ratio of PPC value between M3D and 2D design. We observe that for the adoption of gate-level M3D integration, M3D wafer yield needs to be higher than 90% of 2D wafer yield, and the device manufacturing cost of M3D design should be limited by less than 33% of 2D device manufacturing cost.

Moreover, the experiment result show that FEOL-dominant circuit type has more room for the adoption of M3D, and benefits more from M3D integration than BEOL-dominant circuit type in terms of PPC. This is because the impact of metal stack optimization and giving more routing resources to BEOL-dominant type circuit drastically reduce both power and die area of 2D design compatible to M3D counterpart. The differences in total power and die area between LDPC 2D reference (M8 + Low-K design) and M3D design with the best PPC (M5 / M7 + Low-K design) are only 2% and 25%. However, since the die area of FEOL-dominant circuit type is determined by placement utilization, 50% of footprint saving from M3D technology is guaranteed, resulting in more spaces in terms of die cost for adoption of M3D technology.

Two benchmarks for the previous experiments, DES3 and LDPC, are logic circuits where the number of standard cells in the full-chip 2D design is less than 60k based on

foundry-grade 7nm bulk FinFET. The chip area of these two small circuits is less than $0.01mm^2$. Since the 2D die yield of those extremely small benchmarks is already sufficient, it explains why the huge footprint saving and die cost benefit from M3D technology does not show up. Therefore, the impact of die area of logic-only design on the die cost of M3D and 2D design is evaluated based on the cost models proposed in Section 3.1. Since the 2D die area of BEOL-dominant circuit is effectively reduced when more routing resources are used, footprint saving of gate-level 2-tier M3D design is only 25% as shown in Section 3.3.3. When the die area is determined by placement utilization like FEOL-dominant circuit, 50% of M3D area saving is guaranteed as analyzed in Section 3.3.3.

We assume that the ratio of the die area of 2D design and that of M3D design is fixed in each circuit type, and calculate die cost for each design scheme considering die yield. Figure 3.4 shows that M3D die cost becomes cheaper than 2D die cost along with the increase in die size. M3D design of FEOL-dominant circuit has significant die cost saving compared to 2D design starting from $2mm^2$ while M3D design of BEOL-dominant circuit becomes cheaper from $70mm^2$ as well. In addition, with the same die size of design for two circuit types, the gap for the ratio between 2D and M3D die cost of FEOL-dominant and BEOL-dominant circuit becomes wider along with die size increase. Assuming $100mm^2$ of 2D die size, FEOL-dominant circuit has 2.5 times more cost competitiveness from M3D technology than BEOL-dominant circuit. The result indicates FEOL-dominant circuit benefits sooner and more from M3D technology in terms of cost than BEOL-dominant circuit.

3.5 Summary

This paper studies power, performance, and cost (PPC) tradeoffs with full-chip GDS based cost modeling for 2-tier, gate-level, full-chip GDS monolithic 3D ICs (M3D) built using a foundry-grade 7nm bulk FinFET technology. We propose normalized wafer and die cost models based on the number of metal stacks and die area for 2D and M3D. In our PPC tradeoff study with the simple but self-contained cost models, both 2D and M3D designs

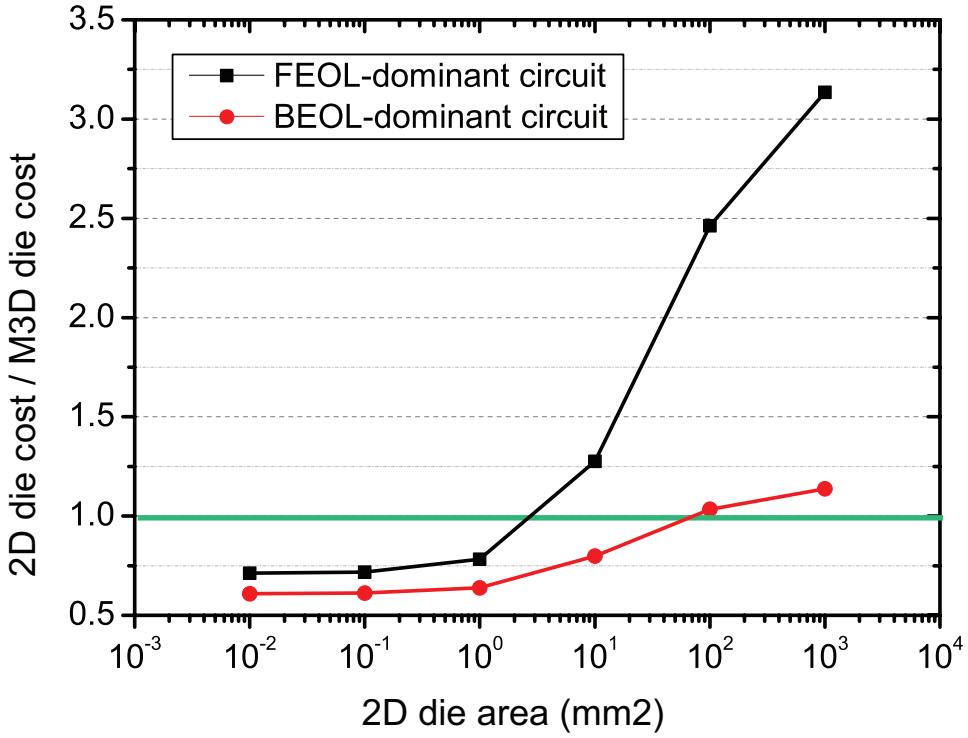

Figure 3.4: Die size impact on the die cost ratio between 2D and M3D. Two different circuit type (FEOL-dominant and BEOL-dominant) are investigated. The region above the green line indicates where the M3D die cost is cheaper than 2D die cost.

are optimized in terms of the number of BEOL metal layers used for routing to obtain the best possible PPC values for the fair comparison. Also, a new CAD methodology for 2-tier gate-level M3D named Projected 2D Flow is developed, that maximizes the placement and routing utilization of M3D design by reducing its footprint by more than 50% compared with that of 2D. Furthermore, this flow allows us to accurately compare RC parasitics of equivalent nets in both 2D and M3D designs since final netlists of these two design flavors are the same.

Based on the experiments with two widely different circuit types (BEOL-dominant vs. FEOL-dominant), it is confirmed that while M3D has indeed a great footprint saving, the PPC quality of M3D is actually worse than that of optimized 2D reference by 34% due to high M3D wafer cost. Our study also shows that, for the adoption of M3D technology at the 7nm era, M3D wafer yield needs to be higher than 90% of 2D wafer yield, and the

2-tier device manufacturing cost of M3D design needs to be limited by less than 33% of 2D device manufacturing cost, and lastly the die area should be large enough ($100mm^2$-scale) to have fruitful die cost reduction from huge M3D footprint saving. Lastly, and counter-intuitively, this study shows that FEOL-dominant type circuit has PPC benefits from M3D technology more and sooner than BEOL-dominant type circuit.

CHAPTER 4
CONCLUSION

Monolithic 3D (M3D) integration has emerged as a viable solution for the massive and silicon-area overhead-free 3D interconnection. However, low thermal budget for top tier fabrication, and the high manufacturing cost are known as the obvious obstacles to adoption of M3D ICs. Although the fabrication process of M3D integration is not fully mature yet, evaluation of the power-performance-cost benefit of M3D ICs considering all these unique challenges should be required at the early phase for the adoption of M3D technology in the 7nm technology node.

This dissertation modeled the impact of low thermal budget top tier fabrication on the device and interconnect integrity, and quantified the degradation of power-performance benefits of M3D ICs built using a foundry-grade 7nm bulk FinFET technology process design kit. A physical design methodology for M3D ICs named Derated 2D is presented to tackle the FEOL/BEOL degradation issues, and experiments showed that proposed design solution offers an efficient timing closure capability to M3D ICs under the various degradation scenario.

The complicated power-performance-cost tradeoffs of M3D ICs were studied based on the highly-accurate, full-chip, GDSII-based wafer and die cost model. A physical design methodology for M3D ICs named Projected 2D is presented to fully optimize the area savings in two-tier M3D ICs, and experiments suggested gate-dominant, large footprint design actually show the power-performance-cost benefit of two-tier gate-level monolithic 3D ICs in the 7nm technology node.

As a future work, merging Derated 2D and Projected 2D flow is pursued. This will provide more thorough analysis to show the power-performance-cost benefit of gate-level M3D ICs under the various scenarios.

REFERENCES

[1] D. Yakimets *et al.*, "Vertical GAAFETs for the Ultimate CMOS Scaling," *IEEE Trans. on Electron Devices*, vol. 62, no. 5, pp. 1433–1439, 2015.

[2] A. Mallik *et al.*, "Maintaining Moore's law: enabling cost-friendly dimensional scaling," vol. 9422, 2015, 94221N–94221N–12.

[3] P. Raghavan *et al.*, "5nm: Has the time for a device change come?" In *Proc. Int. Symp. on Quality Electronic Design*, 2016, pp. 275–277.

[4] A. Mallik *et al.*, "The need for EUV lithography at advanced technology for sustainable wafer cost," vol. 8679, 2013, 86792Y–86792Y–10.

[5] ——, "The economic impact of EUV lithography on critical process modules," in *Proc. SPIE*, vol. 9048, 2014, 90481R–90481R–12.

[6] T. N. Theis and H. S. P. Wong, "The End of Moore's Law: A New Beginning for Information Technology," *Computing in Science Engineering*, vol. 19, no. 2, pp. 41–50, 2017.

[7] P Batude *et al.*, "3-D Sequential Integration: A Key Enabling Technology for Heterogeneous Co-Integration of New Function With CMOS," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 2012.

[8] D. K. Nayak, S. Banna, S. K. Samal, and S. K. Lim, "Power, performance, and cost comparisons of monolithic 3D ICs and TSV-based 3D ICs," in *SOI-3D-Subthreshold Microelectronics Technology Unified Conference*, 2015, pp. 1–2.

[9] M. Vinet *et al.*, "Monolithic 3D integration: A powerful alternative to classical 2D scaling," in *SOI-3D-Subthreshold Microelectronics Technology Unified Conference (S3S)*, 2014, pp. 1–3.

[10] Y.-J. Lee, D. Limbrick, and S. K. Lim, "Power benefit study for ultra-high density transistor-level monolithic 3D ICs," in *Proc. ACM Design Automation Conf.*, 2013, pp. 1–10.

[11] S. Panth, K. Samadi, Y. Du, and S. K. Lim, "Design and CAD methodologies for low power gate-level monolithic 3D ICs," in *Proc. Int. Symp. on Low Power Electronics and Design*, 2014, pp. 171–176.

[12] ——, "Power-performance study of block-level monolithic 3D-ICs considering inter-tier performance variations," in *Proc. ACM Design Automation Conf.*, 2014, pp. 1–6.

[13] P. Batude *et al.*, "3D sequential integration opportunities and technology optimization," in *Proc. IEEE Int. Interconnect Technology Conference*, 2014, pp. 373–376.

[14] F. Luce *et al.*, "Methodology for thermal budget reduction of SPER down to 450C for 3D sequential integration," *Nuclear Instruments and Methods in Physics Research*, vol. 370, pp. 14 –18, 2016.

[15] K. Chang, K. Acharya, S. Sinha, B. Cline, G. Yeric, and S. K. Lim, "Power benefit study of monolithic 3D IC at the 7nm technology node," in *Proc. Int. Symp. on Low Power Electronics and Design*, 2015, pp. 201–206.

[16] S. Panth, K. Samadi, Y. Du, and S. K. Lim, "Placement-Driven Partitioning for Congestion Mitigation in Monolithic 3D IC Designs," *IEEE Trans. on Computer-Aided Design of Int. Circuits and Systems*, vol. 34, no. 4, pp. 540–553, 2015.

[17] X. Dong, J. Zhao, and Y. Xie, "Fabrication Cost Analysis and Cost-Aware Design Space Exploration for 3-D ICs," vol. 29, no. 12, pp. 1959–1972, 2010.

[18] Q. Zou, J. Xie, and Y. Xie, "Cost-driven 3D design optimization with metal layer reduction technique," in *Proc. Int. Symp. on Quality Electronic Design*, 2013, pp. 294–299.